Parametric & Non-Parametric Statistics

⇒ Parametric Statistics :-

: Parametric statistics is defined to have an assumption of normal distribution for its population under study.

- It refers to those statistical techniques which have been developed on the assumption that the data are of certain types.
- In particular, the measure should be an internal scale and the scores should be drawn from a normal distribution (Stratton and Hayes, 1999).

• T-test, F-test & R-test are some of the major parametric statistics used for data analysis.

Assumptions :-

① The population from which the sample have been drawn should be normally distributed.

② The variables involved must have been measured internal or ratio scale.

③ The observation must be independent. The inclusion or exclusion of any case in the sample should not unduly affect the results of study.

④ These populations must have the same variance or, in special cases, must have a known ratio of variance. This we call homosedasticity.

⑤ The samples have equal or nearly equal variances. This condition is known as equality or homogeneity of variances and is particularly important to determine when the samples are small.

⑥ The observations are independent. The selection of one case in the sample is not dependent upon the selection of any other case.

Advantages of Parametric statistics :-

① In the view of the above assumptions, parametric statistics seems to be more reliable and authentic as compared to the non-parametric statistics.

② Such statistics is more powerful to establish the statistical significance of effects and differences among variables.

SHOUT
SHARE
IT

③ It is more appropriate and reliable to use parametric statistics in case of large samples as it promises more accuracy of results.

④ The data to be analysed under parametric statistics are usually from the internal scale.

Disadvantages:-

↳ However, along with many advantages, some disadvantages have also been noted in the parametric statistics.

↳ It is bound to follow the rigid assumption of normal distribution, which further narrows the scope of its usage.

↳ In case of small samples, normal distribution cannot be attained, and thus, parametric statistics cannot be used.

↳ Further, computation in parametric statistics is lengthy and complex because of large samples and numerical calculations

⇒ Non-Parametric Statistics:-

- " " " are not based on the assumption of normal distribution of population. Therefore, these are also known as distribution free statistics.
- They are not bound to be used with internal scale data or normally distributed data. The data with non-continuity are to be tackled with such statistics.
- In the samples, where it is difficult to maintain the assumption of normal distribution, non-parametric statistics are used for analysis.
- The samples with small no. of items are treated with non-parametric statistics because of the absence of normal distribution.
- These can be used even for nominal data along with the ordinal data.
- Some of the usual non-parametric statistics include chi-square, Spearman's rank difference method of correlation, Kendall's rank difference method and Mann-whitney U-test.

~~Assam~~

Assumptions of Non-Parametric Statistics :-

→ We face many situations, where we cannot meet the assumptions and conditions, and thus, cannot use parametric statistical procedures. In such situations, we are bound to apply non-parametric statistics.

→ If our sample is in the form of nominal or ordinal scale, and the distribution of sample is not normally distributed, and also the sample size is very small; it is advisable to make use of the non parametric tests for comparing samples and to make inferences or test the significance or trustworthiness of the computed statistics.

→ In other words, the use of non-parametric tests is recommended in the following situations:-

(i) When sample size is quite small. If the size of the sample is as small as $N=5$ or $N=6$, the only alternative is to make use of non-parametric tests.

(ii) When assumption like normality of the distribution of scores in the population are doubtful, we use non-parametric tests.

(iii) When the measurement of data is available either in the form of ordinal or nominal scales or when the data can be expressed in the form of ranks or in the shape of + signs or - signs and classification like "good-bad", etc. we use non-parametric statistics.

(iv) The nature of the population from which samples are drawn is not known to be normal.

(v) The variables are expressed in the nominal form.

(vi) The data are measures, which are ranked or expressed in numerical scores, having the strength of ranks.

Advantages of Non-parametric statistics :-

- Non-parametric methods can be used to analyse categorical (nominal scaling) data, rank (ordinal scaling) data and interval (ratio scaling) data.
- Non-parametric methods are generally easy to apply and quick to compute when sample size is small.
- Non-parametric methods require few assumptions but are very useful when the scale of measurement is weaker than required for parametric methods. Hence, these methods are widely used and yield a more general broad-based conclusion.

- Non-parametric methods provide an approximate solution to an exact problem whereas parametric methods provide an exact solution to an approximate problem.
- Non-parametric methods provide solution to problems that do not require to make the assumption that a population is distributed normally or any specific shape.

Disadvantages:- Few major disadvantages are as follows:-

- Non-parametric methods should not be used when all the assumptions of the parametric methods can be met. However, they are equally powerful when assumptions are met, when assumptions are not met they may be more powerful.
- Non-parametric methods require more manual computational time when sample size gets larger.
- Table values for non-parametric statistics are not as readily available as of parametric methods.
- Non-parametric tests are usually not as widely used and not well known as parametric tests.

⇒ Use of non-parametric tests:-

In order to achieve the correct results from the statistical analysis, one should know the situations in which the application of non-parametric tests is appropriate. The main reasons to apply the non-n test include the following:-

1. The underlying data do not meet the assumptions about the population sample:-

Generally, the application of parametric tests require various assumptions to be satisfied. For example, the data should follow a normal distribution and the population variance is homogeneous. However, some data samples may show skewed distributions.

The skewness makes the parametric test less ~~no~~powerful because the mean is no longer the best measure of central tendency because it is strongly affected by the extreme values. At the same time, non-parametric tests work well with skewed distributions and distributions that are better represented by the median.

2. The population size is too small :- The sample size is an important assumption in selecting the appropriate statistical method. If a sample size is reasonably large, the applicable parametric test can be used. However, if a sample size is too small, it is possible that you may not be able to validate the distribution of the data. Thus, the application of non-parametric tests is the only suitable option.

3. The analyzed data is ordinal or nominal :- Unlike parametric tests that can work only with the continuous data, non-parametric tests can be applied to other data types such as ordinal or nominal data. For such types of variables, the non-parametric tests are the only appropriate solution.

⇒ Rational for using non-parametric statistics :- Quantities such as means, standard deviations and proportions are all important values and are called 'parameters' when we are talking about the population. Since we usually cannot get data from the whole population, we cannot know the values of the parameters for that population. We can, however, calculate estimates of these quantities for our sample. When they are calculated from sample data, these quantities are called "statistics". A statistic estimates a parameter.

Parametric statistical procedures rely on assumptions about the shape of the distribution (i.e., assume a normal distribution) in the underlying population and about the form or parameters (i.e., means and standard deviations) of the assumed distribution.

While in many situations parametric assumptions are reasonable, we often have no prior knowledge of the underlying distributions. In such situations, the use of parametric statistics can give ~~wrong~~ mis-leading or even wrong results.

In such situations, we need statistical procedures which are insensitive to the model assumptions in the sense that the procedures retain their properties in the neighbourhood of the model assumptions.

Non-parametric statistical procedures rely on few or no assumptions about the shape of parameters of the population distribution from which the sample is drawn. To put it in another way, non-parametric means that the standard methods cannot be used. Non-parametric tests are commonly called distribution free or rank order tests.

Parametric Test	Non-Parametric Test
→ Make numerous or stringent assumptions about parameters	→ Donot make — — — — — —
→ A parametric test focuses on the mean difference	→ Equivalent non-parametric test focuses on the difference b/w medians.
→ If the mean accurately represents the center of distribution and sample size is large enough, consider a parametric test because they are more powerful.	→ If the mean better represents the center of distribution, consider the non-parametric test even in case of a large sample.
→ Original data score is used.	→ Data need to be changed from scores to rank or signs
→ Powerful	→ less powerful
→ Efficient	→ less efficient
→ Mean & SD are important	→ mean & SD are not important.
→ Sample size more than 30	→ Normally sample size less than 30.
→ Information about the population is completely known	→ No information about the population.
→ Data should be normally distributed	→ distribution free tests
→ Applicable only for variable	→ Applied to both variable & attributes
→ Not applicable for nominal scale data	→ Exist for nominal and ordinal scale
→ Null hypothesis is made on Parameters of the population distribution.	→ The null hypothesis is free from parameters
→ More "conservative"	→ More "liberal".
→ Less likely to make a Type-I error	→ More likely to make a Type-I error.

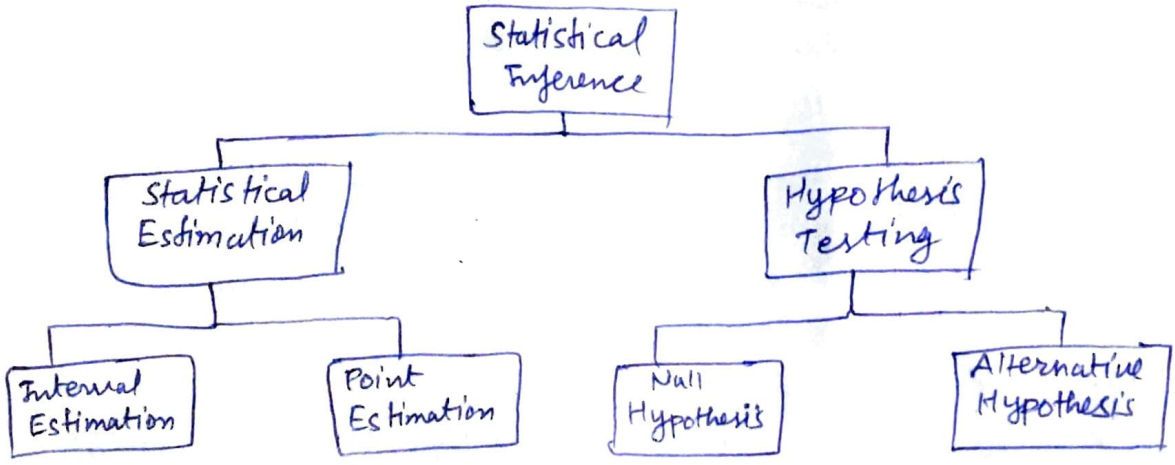
⇒ Inferential Statistics:-

Statistical inference is that branch of statistics, which is concerned with using probability concept to deal with uncertainity in decision-making. The field of statistical inference has had a fruitful development since the latter half of the 19th century.

It refers to the process of selecting and using a sample statistic to draw inference about a population parameter based on a subset of it, i.e. the sample drawn from the population. Statistical Inference treats two different classes of problems:

- Hypothesis testing, i.e. to test some hypothesis about parent population from which the sample is drawn.
- Estimation, i.e. to use the 'statistics' obtained from the sample as estimate of the unknown 'parameter' of the population from which the sample is drawn.

In both these cases, the particular problem at hand is structured in such a way that inferences about relevant population values can be made from sample data.



When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inferences for generalisations about the population from the information embodied in the sample. Statistical estimation, or briefly estimation, is concerned with the methods by which population characteristics are estimated from sample information. It may be pointed out that the true value of parameter is an unknown constant that can be correctly ascertained only by an exhaustive study of the population. However, it is ordinarily too expensive or it is infeasible to enumerate complete population to obtain the required information. In case of finite population, the cost complete censuses may be prohibitive and in case of

infinite population, complete enumerations are impossible. A realistic objective may be to obtain a guess of estimate of the unknown true value or an interval of plausible values from the sample data and also to determine the accuracy of the procedure. Statistical estimation procedures provide us with the means of obtaining estimates of population parameters with desired degrees of precision.

Characteristics of Estimators:- The basic characteristics of estimates are as follows:-

- Unbiasedness and Minimum Variance:- A statistic T is defined to be unbiased for a parameter θ if expectation of T is i.e. $E(T) = \theta$. On the other hand, if, $E(T) = \theta + a(\theta)$, then the difference $a(\theta) = E(T) - \theta$ is known as bias. The bias is known to be positive if $a(\theta) > 0$ and negative if $a(\theta) < 0$. Our first priority would be to select an unbiased estimator of θ. However, there may be many unbiased estimators of θ. If $x_1, x_2, \ldots x_n$ denote n sample observations from a population with an unknown parameter θ, then any of the n observations or any linear function of them would be an unbiased estimator of θ.
- Consistency:- If T is an estimator of θ, then it is obvious that T should be in the neighbourhood of θ. T is known to be consistent for θ, if the difference between T and θ can be made as small as we please by increasing the sample size n sufficiently.
- Efficiency:- A statistic T is called as an efficient estimator of θ if it has the minimum standard error among all the estimators of θ for a fixed sample size n. Both the sample mean and sample median are consistent estimators for μ. But standard error of sample mean is less than that of sample median. Hence, sample median is only a consistent estimator of μ, whereas sample mean is both consistent and efficient estimator of μ.
- Sufficiency:- A statistic T is known to be a sufficient estimator of θ if T contains sufficient information about θ so that we do not have to look for any other estimator of θ. Sample mean ($\bar{x}$) is a sufficient estimator of μ.

With respect to estimating a parameter, the following two types of estimates are possible:-

(1) Point Estimates:- A point estimate is a single number computed from the sample values $x_1, x_2, \ldots x_n$ which is used as an estimate of the unknown population parameter. The procedure in point estimation is to select a random sample of n observations, $x_1, x_2, \ldots x_n$ from a population $f(x, \theta^*)$ and then to use some preconceived method to arrive from these observations at a number say $\hat{\theta}$ (read theta hat) which we accept as

an estimator of θ. The estimator $\hat{\theta}$ is a single point on the real number scale, and thus, the name point estimate. $\hat{\theta}$ depends on the random variable that generate the sample, and hence, it too is a random variable with its own sampling distribution.

(2) Interval Estimates:-

As distinguished from a point estimate, which provides one single value of the parameter, an interval estimate of a population parameter is a statement of two values between which it is estimated that the parameter lies. An interval estimate would always be specified by two values, i.e. the lower one and the upper one. In more technical terms, interval estimation refers to the estimation of a parameter by a random interval, called the confidence interval, whose end points L & U with $L < U$ are function of the observed random variables such that the probability that the inequality $L < \theta < U$ is satisfied in terms of pre-determined number, $1-\alpha$. L and U are called the confidence limits and are the random end points of interval estimate.

Since in an interval estimate, we determine an interval of plausible values, hence the name interval estimation. Thus, on the basis of sample study, if we estimate the average income of the people living in a village as ₹875, it will be a point estimate on the other hand, if we say that the average income could lie between ₹800 and ₹950, it will be an interval estimate.

On comparing these two methods of estimation, we find that point estimation has an advantage as much as it provides an exact value for the parameter under investigation. This merit, however, is also the defect of a point estimate. Being a single point on the real number scale, a point estimate doesnot tell us how close the estimator is to the parameter being estimated. Moreover, in scientific investigation, it is usually not necessary to know exact value of a parameter - rather it is desirable to have some degree of the confidence that the value obtained is within a certain range. The interval estimate does provide such confidence, and hence, interval estimate should generally be employed in practice.

⇒ Differentiate between Inferential & Descriptive statistics.

Descriptive Statistics	Inferential Statistics
→ Summarize essential features of data	→ Estimates, predictions, forecasts, generalizations.
→ It concerned with describing the population under study	→ It focuses on drawing conclusions about population, on basis of sample analysis and observation.
→ organize, analyze and present data in a meaningful way.	→ compares, test and predicts data.
→ Form of final Results charts, Graphs and Tables	→ Form of final result - Probability
→ Used to describe a situation.	→ Used to explain the chances of occurence of an event.
→ Deals with central tendency and spread of the frequency distribution.	→ More details such as hypothesis tests and confidence interval are studied.
→ The measures of descriptive statistics (mean, median, mode) are numbers.	→ The measures in inferential statistics are not always exact numbers.
→ The conclusions cannot be made beyond the given data	→ The educated predictions & guesses can be made on the basis of the parameters of the given population.
→ Deals with small samples which enables us to produce results without errors.	→ Inferential statistics takes whole population for drawing conclusions which may not have the extent of required accuracy.

Descriptive statistics :-

" " are used to describe the basic features of the data in a study. Descriptive statistics help us to simplify large amounts of data by organizing and summarizing it in a sensible way so that useful information is produced. They form the basis of virtually every quantitative analysis of data.

Descriptive statistics do not allow us to make conclusions beyond the data we have analyzed or reach conclusions. Still descriptive statistics are very imperative because if we simply presented our raw data, it would be difficult to visualize what the data was showing, especially if there was a lot of it.

There are two classes of descriptive statistics, one class includes organizing data with the help of tables and graphs and the other class includes summarizing data with the help of numerical measures and index numbers.

Advantages of Descriptive Statistics :-

- It is essential for arranging and displaying data.
- It forms the basis of rigorous data analysis
- It is easier to work with, interpret, and discuss than raw data.
- It helps in examining the tendencies, variability, and normality of a data set.
- It can be rendered both graphically & numerically.
- It forms the basis for more advanced statistical methods.

Disadvantages of Descriptive Statistics :-

- It can be misused, misinterpreted, and incomplete.
- It can be of limited use when samples and population are small.
- It offers little information about causes and effects.
- It can be dangerous if not analysed completely.
- There is a risk of distorting the original data or losing important detail.

⇒ Levels of measurement and indicate the data for which they are used:-

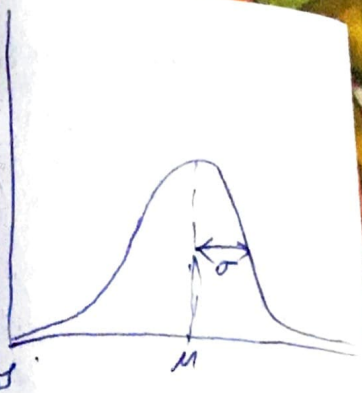
When deciding which statistical test to use, it is important to identify the level of measurement associated with the dependent variable of interest.

There are four main types:-

- Nominal → A nominal level variable has categories only. There is no dimension or numbering to the categories. The numbers assigned to each category are MEANINGLESS and are just a grouping procedure.
  E.g:- Gender, Eye Colour, Types of building.
- Ordinal → The numbers have an order! The order is meaningful. The numbers mean something. With ordinal scales, the order of the values is important and significant, but the differences between each is not really known. We have an order but they are not real numbers.
  E.g:- Likert scales (Strongly Agree - Agree - Neutral - Disagree - Strongly Disagree)
- Interval:- An interval variable has categories that are 'real numbers' with common units of measurement, but there is NO TRUE ZERO. The zero does not indicate 'a complete absence' of something.
  E.g:- Date, Temperature, length, weight, frequency
- Ratio:- A ratio level variable is the same as an interval variable, but they have a 'true zero'. You have real numbers with real intervals and a zero means 'a lack of' or an 'absence of' whatever you are measuring.
  E.g:- Age, sales figure, income, GPA, years of experience.

Normal curve :-

A normal curve is a bell-shaped curve which shows the probability distribution of a continuous random variable.



Moreover, the normal curve represents a normal distribution. The total area under the normal curve logically represents the sum of all probabilities for a random variable. Hence, the area under the normal curve is one. Also, the standard normal curve represents a normal curve with mean 0 and standard deviation 1. Thus, Parameters involved in a normal distribution is mean (μ) and std. deviation (σ).

Properties of the normal probability curve :-

→ Symmetrical (That is, when you cut it in half the two pieces are mirror images of each other).

→ Unimodel (Its distribution has a bump in the middle, with tails going down and out to the left and right).

→ Asymptotic to the X-axis (The normal curve approaches, but never touches, the x-axis as it extends farther and farther away from the mean)

→ Mean = Median = Mode (The mean, median and mode are the same and lie directly in the middle of the distribution)

→ Area under curve equals 1.

→ The total Area under Normal curve may be also considered 100 percent probability.

→ The maximum ordinate occurs at the center.

→ The normal curve is a mathematical model in behavioural Sciences specially in Mental Measurement.

Divergence in Normal Distribution:- In normal distribution, most cases fall in the middle but there are cases in which central tendency do not exhibit normal behaviour.

Statistically, two numerical measures of shape - Skewness and kurtosis can be used to test for normality. If either of these values is not close to zero, then a data set is not normally distributed.

Causes for Diversion:-

→ If sample size is small or sample is biased one, skewness is possible in the distribution of scores.

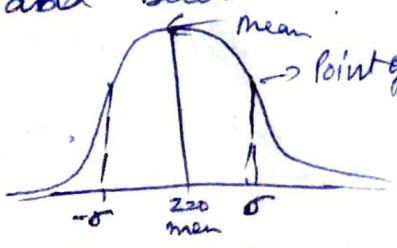
→ If the measuring tool or test is inappropriate for the group on which it has been administered, or poorly made, the asymmetry, is likely to occur in the normal distribution of scores.

→ A real lack of normality in the trait that is measured, produces skewness and kurtosis; e.g., teacher's salaries in a school, where most of the teachers are experienced and are close to the maximum on the salary schedule, will give a negative skewed distribution.

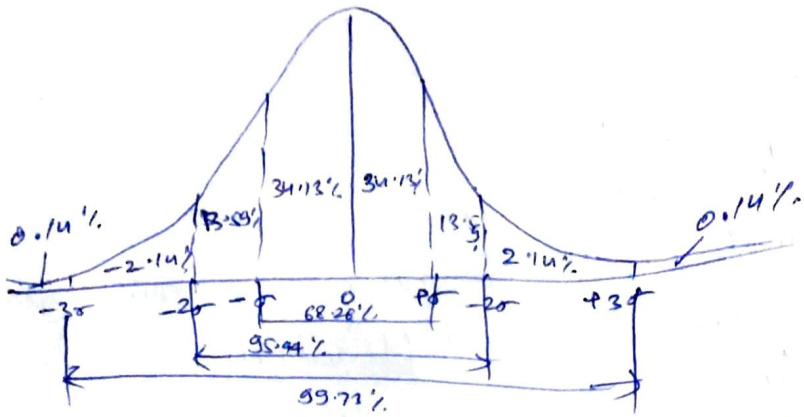
→ A poorly constructed test may cause asymmetry in the normal distribution of the scores.

Properties contd:-

→ **The point of influx occur at point ± 1 stand. deviation (± 1σ):-** The normal curve changes its direction from convex to concave at a point recognised as point of influx. If we draw the perpendiculars from these two points of influx of the curve to the horizontal axis; touch at a distance one std. deviation unit from above and below the meand central point)



→ The total percentage of area of the normal curve between two points of inflexation is fixed i.e. Approximately 68.26%. area of the curve lies within the limits of ± 1 Std. deviation unit from the mean.



→ The normal curve is bilateral :- 50% area of the curve lies to the left side of the maximum central ordinate and 50% of the area lies to the right side. Hence curve is bilateral.

Uses of Normal Distribution Curve :-

→ To determine the percentage of cases (in a normal distribution) within given limits or scores.

→ To determine the %age of cases that are above or below a given score or reference point.

→ To determine the limits of scores which include a given % of cases to determine the %ile rank of an individual or a student in his own group.

→ To find out the %ile value of an individual on the b of his percentile rank.

→ Dividing a group into sub-groups according to ce ability and assigning the grades.

→ To compare the two distributions in terms of overlapping.

→ To determine the relative difficulty of test items

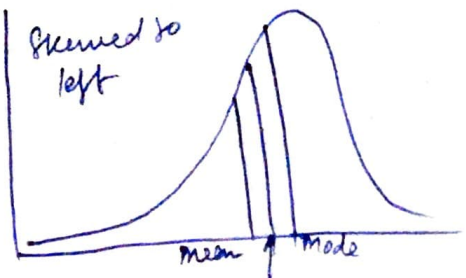
Interpretation of Normal Curve / Normal Distribution

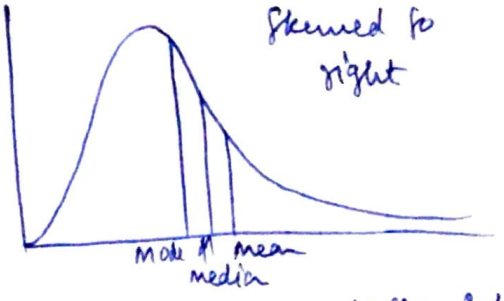
(1) The measured trait is normally distributed in the universe.
(2) Most of the cases i.e. individuals are average in the measured trait and their %age in the total population is about 68.26%.
(3) Approximately 15.87% (50 - 34.13%) of cases are high in the trait measured.
(4) Similarly, 15.87% of cases are low in the trait measured.
(5) The test which is used to measure the trait is good.
(6) The test which is used to measure the trait has good discrimination power as it differentiates blw poor, average and high ability group individuals.
(7) The items of the test used are fairly distributed in terms of difficulty level.

Meaning Divergence from Normal Curve

① Skewness :- A distribution is said to be 'skewed' when the mean and median fall at difference points in the distribution and the balance i.e. the point of center of gravity is shifted to one side or the other to the left or right. In a normal distribution the mean equals, the median exactly and the skewness is of course zero ($Sk = 0$).
There are two types of skewness which appear in the normal curve

(a) Negative Skewness :- Distribution said to be skewed negatively or to the left when scores are massed at the high end of the scale, i.e. the right side of the curve are spred out more gradually toward the low end i.e. the left side of the curve. In negatively skewed distribution the value of median will be higher than that of the value of the mean.



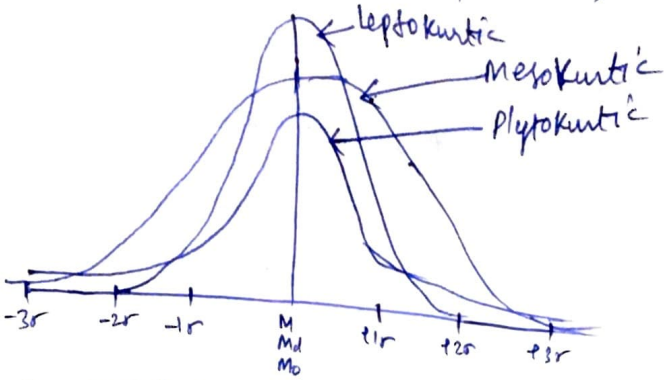


(b) Positive skewness :- Distributions are skewed positively or to the right, when scores are massed at low, i.e. the left end of the scale, and are spread out gradually toward the high or right end as shown in figure.

- o mean -

(2) Kurtosis :- Kurtosis is a measure of the shape of a distribution. It is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak.

The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic



When the distribution and related curve is normal, the value of Kurtosis is 0.263 (KV = 0.263), the shape is mesokurtic. If the value of KV is greater than 0.263, the distribution and related curve obtained will be platykurtic. When the value of KV is less than 0.263, the distribution and related curve obtained will be leptokurtic.

⇒ Two-way Analysis of Variance :-

In two-way analysis of variance, usually the two independent variables are taken simultaneously. It has two main effects and one interactional or joint effect on dependent variable. In such condition, we have to use ANOVA in two ways, i.e. vertically as well as horizontally or we have to use ANOVA, column and row wise. Suppose we are interested to study the intelligence, i.e. I.Q. level of boys and girls studying in VIII class in relation to their level of socio economic status (SES). In such condition, we have the following 3×2 design as shown in the table below.

Groups	Levels of SES			
	High	Average	Low	Total
Boys	M_{HB}	M_{AB}	M_{LB}	M_B
Girls	M_{HG}	M_{AG}	M_{LG}	M_G
Total	M_H	M_A	M_L	M

In the table,
M: Mean of intelligent scores.
M_{HB}, M_{AB} & M_{LB}: Mean of intelligence scores of boys belonging to different levels of SES, i.e. High, Avg & low respectively
M_H, M_A & M_L: Mean of intelligent scores of students belonging to diff. levels of SES respectively

M_{HG}, M_{AG}, M_{LG}:- Mean of intelligent scores of girls belonging to diff. levels of SES respectively

M_B, M_G:- Mean of the intelligence scores of boys & girls respectively.

From the above 3x2 contingency table, it is clear, first we have to study the significant difference in the means column wise or vertically, i.e. to compare the intelligence level of the students belonging to different categories of SES.

Secondly, we have to study the significant difference in the means row wise or horizontally, i.e. to compose the intelligence level of the boys and girls.

Then we have to study the interactional or joint effect of sex and SES & rows. on intelligence level i.e. we've to compare the significant diff. in the cell means of columns

We've more than two groups, and to study the independent as well as interaction effect of the two variables, viz. socio-economic status and sex on dependent variable, viz. intelligence in terms of I.Q., we have to use two-way ANOVA.

Therefore, in two-way analysis of variance technique, the following type of effects are to be tested:

- Significance of the effect of A variable on D.V.
- " " " B " " D.V.
- " " " interaction effect of AxB variables on D.V.

In two-way ANOVA, the format of summary table after applying the ANOVA is as under:

Source of variance	df	SS	MSS	F-Ratio
Among the groups				
Between the grp A	$K_A - 1$	SS_A	MSS_A	$F_1 = \frac{MSS_A}{MSS_W}$
Between the grp B	$K_B - 1$	SS_B	MSS_B	$F_2 = \frac{MSS_B}{MSS_W}$
Interselation AxB	$(K_A-1)(K_B-1)$	SS_{AxB}	MSS_{AxB}	$F_3 = \frac{MSS_{AxB}}{MSS_W}$
Within the groups (Error Variance)	$N - K_A - K_B$			
Total	$N-1$			

For interpretation of the obtained F-ratios, we have to evaluate each F-ratio value with the F-ratio given in F-table, keeping in view the corresponding greater and smaller df and the level of confidence. There may be two possibilities.

→ All the obtained F-ratios may be found insignificant even at 0.05 level. This shows that there is no independent (i.e. individuals) as well as interaction (i.e. joint) effect of the two independent variables on dependent variable. Hence, null hypothesis will retain. There is no need to do further calculations.

→ All the three obtained F-ratio's may be found significant either at 0.05 level of significance or at 0.01 level of significance. This shows that there is a significant independent (i.e. individual) as well as interactional (i.e. joint) effect of the independent variables on the dependent variable. Therefore, the null hypothesis is rejected.

Merits and Demerits of 2-way ANOVA :-

Merits:-

→ This technique is used to analyse two types of effects, viz. main effects and interaction effects.

→ More than two factors effects are analysed by this technique

→ For analysing the data obtained on the basis of factorial designs, this technique is used.

→ This technique is used to analyse the data for complex experimental studies.

Demerits:-

→ When there are more than two classifications of a factor(s) of study, F-ratio value provides global picture of difference among the main treatment effects. The inference can be specified by using t-test in case when F-ratio is found significant for a treatment.

→ This technique also follows the assumptions on which one-way ANOVA is based. If these assumptions are not fulfilled, the use of this technique may give us spurious results.

→ This technique is difficult and time consuming.

→ As the number of factors are increased in a study, the complexity of analysis is increased and interpretation of results become difficult.

→ This technique requires high level arithmetical and calculative ability. Similarly, it also requires high level of imaginative and logical ability to interpret the obtained results.

⇒ **F-ratio test and t-ratio test are complementary to each other**

↳ F-ratio test and t-ratio test are complementary to each other, because 't' is followed when 'F' value is significant for the specification of inferences.

↳ F-test is followed, when t-value is not found significant. Because within groups, variance is not evaluated by t-test. It evaluates only the difference between variance.

↳ There is a fixed relation between 't' and 'F'. The 'F' is the square of 't', while 't' is a square root of 'F".

↳ $F = t^2$ or $t = \sqrt{F}$

ANOVA :- In its simplest form, the ANOVA can be used when two or more than two groups are compared on the basic of certain traits or characteristics or different treatments of simple independent variable is studied on a dependent variable and having two or more than two groups.

It is to be noted here that when we have taken a large group or a finite population, to represent its total units the symbol 'N' will be used.

When the large group is divided into two or more than two subgroups having equal no. of units, the symbol 'n' will be used and for no. of subgroups the symbol 'k' will be used.

Merits and limitations of ANOVA :-

⇒ Merits :-

(1) It is the improved technique over the t-test or z-test. It evaluates both types of variances, i.e. 'between' & 'within'.

(2) This technique is used for ascertaining the difference among several groups or treatments at a time. It is an economical device.

(3) It can involve more than one variable in studying their main effects and interaction effects.

(4) In some of the experimental design, e.g. simple random design and levels x-treatment designs are based on one-way ANOVA

(5) If 'F' is not significant, f-test must be followed to analyse the difference b/w two means.

⇒ Limitations :-

(1) The analysis of variance technique is based on certain assumptions, eg. normality and homogeneity of the variances among the groups. The departure of the data from these assumptions may affect adversely on the inferences.

(2) The F-value provides global findings of difference among grps, but it cannot specify the difference. Therefore, complete ANOVA, the t-test is followed for specifying the statistical inference.

(3) It is a time-consuming process and requires the knowledge & skills of arithmetical operations as well the high vision for interpretation of results.

(4) For the use of F-test, the statistical table of F-value is essential, because without it, results cannot be interpreted.

Hypothesis Testing - A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. It refers to the formal procedures used by statisticians to accept or reject statistical hypotheses, using data measured in a sample.

Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed. Hypothesis testing is an essential procedure in statistics that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population. Hence, it is also referred to as statistical decision-making process.

In hypothesis testing, we are interested in testing two mutually exclusive hypotheses, called the null hypothesis (denoted H_0) and the alternative hypothesis (denoted H_1 or H_a). The null hypothesis is the statement being tested. Usually, the null hypothesis is a statement of 'no effect', 'no difference', or 'no association' whereas the alternative hypothesis is a statement of 'effect', 'difference', or 'association'. So, both the hypothesis is opposite to each other hence, when one of these hypothesis is false, the other must be true. Null hypothesis is assumed to be true until proven otherwise. However, alternative is the hypothesis the researcher hopes to bolster.

It is important to note that both the null and alternative hypotheses should be stated together and before any statistical test of significance is conducted. In other words, we, technically, are not supposed to do the data analysis first and then decide on the hypotheses afterwards.

Errors in Hypothesis Testing :-

No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a chance of drawing an incorrect conclusion. It is not about making errors in calculations or even about using the wrong procedures. These errors are situations in which the right procedures lead to the wrong decisions. When we do a hypothesis test, two types of errors are possible : type I and type II errors

→ Type I error :- when the null hypothesis is true and we reject it, we make a type I error. The probability of making a type I error is α, which is the level of significance we set for our hypothesis test. An α of 0.05 indicates that we are willing to accept a 5% chance that we are wrong when we reject the null hypothesis.

→ Type - II error :- when the null hypothesis is false and we fail to reject it, we make a type II error. The probability of making a type II error is β, which depends on the power of the test.

Null Hypothesis :- (H_0) is a statistical hypothesis that states that there is NO difference or No relationship b/w a parameter and a specific value. It represents the status quo. It is the statement about the population parameter that is assumed to be true unless there is convincing evidence to the contrary. It is the hypothesis the researcher is trying to disprove.

Examples of the Null Hypothesis :- To write a null hypothesis, first start by asking a question. Rephrase that question in a form that assumes no relationship b/w the variables. In other words, assume a treatment has no effect.

Question → Are teens better at math than adults?

Null Hypo :- Age has no effect on mathematical ability.

Ques.1 Degree of Relationship b/w the 2 data given below using Spearman's Rho:

	A	B	C	D	E	F	G	H	I	J
Data 1:	22	23	29	20	25	27	30	34	37	35
Data 2:	35	39	22	40	31	45	30	28	25	20

Sol Let, Data 1 be 'X' and Data 2 be 'Y'

X	Y	Rank Rx	Rank Ry	D = Rx - Ry	D^2
22	35	2	7	-5	25
23	39	3	8	-5	25
29	22	6	2	4	16
20	40	1	9	-8	64
25	31	4	6	-2	04
27	45	5	10	-5	25
30	30	7	5	2	04
34	28	8	4	4	16
37	25	10	3	7	49
35	20	9	1	8	64
		55	55	0	292

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6 \times 292}{10 \times 99} = 1 - \frac{292}{165} = 1 - 1.77$$

$$= -0.77$$

The value of Spearman correlation coefficient b/w the given sets of scores is -0.77, hence it is concluded that there is a strong negative correlation b/w the given sets of data.

ANOVA

Ques: 2 A research was carried out to find the effectiveness of three techniques of stress management. The data collected from three different groups on stress scale are given below. Find out using ANOVA if the obtained differences are significant:

Group A (Training 1): 2, 4, 5, 6, 7
Group B (Training 2): 3, 2, 3, 2, 4
Group C (Training 3): 3, 6, 2, 4, 7

Critical value: 0.05 level of significance = 19.41
0.01 level of significance = 99.46

Ans $H_0: M_1 = M_2 = M_3$: There is no statistical significant difference exists in the effectiveness of three techniques of stress mgmt.

H_1: There is a statistical significant difference exists in the effectiveness of three techniques of stress mgmt.

In this example, $k = 3$ (i.e. three techniques of stress mgmt)
$n = 5$ (i.e. each group is having 5 observations)
$N = 15$ (i.e. the total no of observations)

Here,
$df_{between} = k-1 = 3-1 = 2$
$df_{within} = N-k = 15-3 = 12$
$df_{total} = N-1 = 15-1 = 14$

S.No.	Group A		Group B		Group C	
	x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
1	2	4	3	9	3	9
2	4	16	2	4	6	36
3	5	25	3	9	2	4
4	6	36	2	4	4	16
5	7	49	4	16	7	49
Σx	24	130	14	42	22	114
$(\Sigma x)^2$	576		196		484	
$\Sigma(\Sigma x)$	$24+14+22 = 60$					
$\Sigma(\Sigma x^2)$	$576+196+484 = 1256$					

$$SS_B = \sum \frac{(\Sigma x)^2}{n} - \frac{[\Sigma(\Sigma x)]^2}{N}$$

$$= \left(\frac{24^2}{5} + \frac{14^2}{5} + \frac{22^2}{5}\right) - \left(\frac{24+14+22}{15}\right)^2$$

$$= 11.2$$

$$SS_W = \left(\Sigma(x_1^2) - \frac{(\Sigma x_1)^2}{n}\right) + \left(\Sigma(x_2^2) - \frac{(\Sigma x_2)^2}{n}\right) + \left(\Sigma(x_3^2) - \frac{(\Sigma x_3)^2}{n}\right)$$

$$= \left(130 - \frac{24^2}{5} + 42 - \frac{14^2}{5} + 114 - \frac{22^2}{5}\right)$$

$$= 34.8$$

$$MS_B = \frac{SS_B}{df_B} = \frac{11.2}{2} = 5.6$$

$$MS_W = \frac{SS_W}{df_W} = \frac{34.8}{12} = 2.9$$

$$F\text{-Ratio} = \frac{MS_B}{MS_W} = \frac{5.6}{2.9} = 1.93$$

~~Source of Vo~~

Source of variation	Sum of squares	Degree of freedom	Mean square	F Ratio
Between sample	11.2	2	5.6	1.93
Within sample	37.8	12	2.9	
	46	14		

Critical values of F = 99.46 at 0.01 level of significance

Critical values of F = 19.41 at 0.05 level of significance

Since it is observed that F(calculated) < F(Critical), at 0.01 and 0.05 level of significance, it is then concluded that the null hypothesis is accepted.

Therefore, there is enough evidence to claim that there is no statistical significant difference exists in the effectiveness of three techniques of stress management for all three different groups on stress scale, at $\alpha = 0.05$ & 0.01 level of significance.

Q. compute **Mann-Whitney U test** with the help of following data:-

Data 1: 10, 24, 14, 15, 30, 17, 29

Data 2: 20, 12, 16, 18, 36, 38, 50

Data 1	Rank	Data 2	Rank
10	1	20	8
24	9	12	2
14	3	16	5
15	4	18	7
30	11	36	12
17	6	38	13
29	10	50	14
ΣR_1	44	ΣR_2	61

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U = 7 \times 7 + \frac{7(7+1)}{2} - 44$$

$$U = 49 + 28 - 44 = 33$$

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

$$U' = 7 \times 7 + \frac{7(7+1)}{2} - 61$$

$$U' = 49 + 28 - 61 = 16$$

$\therefore$ **U = 33 & U' = 16**

Q.3 Compute **Kendall's Tau** for the following data:

	A	B	C	D	E
X	4	7	8	9	3
Y	3	4	7	8	9

Ans The data given is as below

	A	B	C	D	E
X	4	7	8	9	3
Y	3	4	7	8	9

Step1 :- Convert the values into the Ranks (Lowest value will get Rank 1)

	A	B	C	D	E
R_X	2	3	4	5	1
R_Y	1	2	3	4	5

R_X is Rank of X variable & R_Y is Rank of Y variable.

Step2 :- Arrange R_X into the ascending order & arrange R_Y accordingly

R_X	1	2	3	4	5
R_Y	5	1	2	3	4

Step3 :- Compute S

R_Y	5	1	2	3	4	Total
	5	−	−	−	−	− 4
		1	+	+	+	+ 3
			2	+	+	+ 2
				3	+	+ 1
					4	0
Grand Total (S)						+2

$$\tau = \frac{2S}{[n(n-1)]} = \frac{2\times 2}{5(5-1)} = \frac{4}{20} = \underline{\underline{0.20}} \text{ Ans}$$

Ques.4 Find out if relationship exists between the data given below with the help of **Pearson's Product Moment coefficient of correlation.**

	A	B	C	D	E	F	G	H	I	J
Data 1:	2	3	4	7	8	9	2	3	4	8
Data 2:	10	7	8	2	3	1	10	10	7	2

Soln Let, Data 1 be 'x' and Data 2 be 'y'

x	y	x^2	y^2	xy
2	10	4	100	20
3	7	9	49	21
4	8	16	64	32
7	2	49	4	14
8	3	64	9	24
9	1	81	1	9
2	10	4	100	20
3	10	9	100	30
4	7	16	49	28
8	2	64	4	16
Total: 50	60	316	480	214

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}}$$

$$r = \frac{214 - \frac{50 \times 60}{10}}{\sqrt{\left(316 - \frac{(50)^2}{10}\right)\left(480 - \frac{(60)^2}{10}\right)}} = \frac{214 - 300}{\sqrt{(316 - 250)(480 - 360)}}$$

$$r = \frac{-86}{88.95} = -0.966$$

The value of pearson correlation coefficient between the gi
sets of scores is -0.966, hence it is concluded that ther
is a very strong negative correlation between the gi
sets of data.

... the following data:

Q.5 Compute **Regression equation for X and Y** based on the following data:

	A	B	C	D	E
X	2	3	2	5	8
Y	3	2	4	1	10

Sol

X	Y	X^2	Y^2	XY
2	3	4	9	6
3	2	9	4	6
2	4	4	16	8
5	1	25	1	5
8	10	64	100	80
Total	20	20	106	130	105

Here, $n = 5$

$\bar{X} = \frac{\Sigma X}{n} = \frac{20}{5} = 4$

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{20}{5} = 4$

$(\Sigma X)^2 = (\Sigma Y)^2 = 20^2 = 400$

Regression Eqn - 'x' on 'y'

$S_X = \sum X^2 - \frac{(\Sigma X)^2}{n} = 106 - \frac{400}{5} = 26$

$cov_{XY} = \sum XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 105 - \frac{400}{5} = 25$

$b = \frac{cov_{XY}}{S_X} = \frac{25}{26} = 0.96$

$a = \bar{Y} - b\bar{X} = 4 - (0.96 * 4) = 0.16$

$\Rightarrow \boxed{\hat{Y} = a + bX} \Rightarrow \boxed{\hat{Y} = 0.16 + 0.96X}$

Regression eqn - (Y) on 'x'

$S_Y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 130 - \frac{400}{5} = 50$

$cov_{XY} = 25$

$b = \frac{cov_{XY}}{S_Y} = \frac{25}{50} = 0.5$

$a = \bar{X} - b\bar{Y} = 4 - (0.5 * 4) = 2$

$\hat{X} = a + bY \Rightarrow \boxed{\hat{X} = 2 + 0.5Y}$

Chi-Square

Ques.6 For question "How often do you exercise"? the replies given by males and females were categorized as frequently, occasionally, rarely and never. Is there any association between gender & frequency?

	Frequently	Occasionally	Rarely	Never
Males	10	5	4	6
Females	20	10	3	2

Critical Value:- for 0.01 level of significance = 11.345
for 0.05 level of significance = 7.815

Ans Null Hypothesis:- There is no association b/w the gender and frequency for the replies for the question, "How often do you exercise?"

Answers Given		Gender: Males	Gender: Females	Total	X^2
Frequently	Observed value (O)	10	20	30	
	Expected value (E)	12.50	17.50		
	O − E	−2.50	2.50		
	$(O-E)^2$	6.25	6.25		
	$(O-E)^2/E$	0.50	0.36		0.86
Occasionally	O	5	10	15	
	E	6.25	8.75		
	O − E	−1.25	1.25		
	$(O-E)^2$	1.56	1.56		
	$(O-E)^2/E$	0.25	0.18		0.43
Rarely	O	4	3	7	
	E	2.92	4.08		
	O − E	1.08	−1.08		
	$(O-E)^2$	1.17	1.17		
	$(O-E)^2/E$	0.40	0.29		0.69
Never	O	6	2	8	
	E	3.33	4.67		
	(O − E)	2.67	−2.67		
	$(O-E)^2$	7.11	7.11		
	$(O-E)^2/E$	2.13	1.52		3.66
	Total	25	35	60	5.63

Here, Expected value is calculated as

$$\text{Expected value} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total}}$$

For every cell, Expected value for the observed value will be calculated with the help of above formula. For example, For 'Frequently', 'Males' the observed value is 10, so the expected value for the same will be

$$\text{Expected value} = \frac{25 \times 30}{60} = \frac{750}{60} = 12.5$$

Likewise, for all the given 8 observed values in the example, we need to calculate the expected value by taking respective row and column total into consideration. The final expected values are given in table.

Now, chi-Square:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 5.63$$

Whereas, At 0.01 level of significance, calculated chi-Square Value (5.63) is less than critical chi-square table value (11.345), so Null hypothesis is accepted.

whereas, at 0.05 level of significance, calculated value of chi-Square (5.63) is less than critical chi-square table value (7.815), so Null hypothesis is also accepted.

Hence, we can say that there is no association b/w the gender and the frequency for the replies for the question, 'How often do you exercise?' at 0.01 and 0.05 level of significance.

Ques:7 Compute 't' test for the following data given below.

Group A: 10, 4, 3, 2, 4, 2, 5, 10, 5, 5

Group B: 4, 6, 8, 2, 9, 1, 12, 13, 10, 10

Critical value: 2.10 at 0.05 level of significance, 2.88 at 0.01 L.o.S.

Ans

Group A (x)	x - Mean(M)	$(x-M)^2$
10	5	25
4	-1	1
3	-2	4
2	-3	9
4	-1	1
2	-3	9
5	0	0
10	5	25
5	0	0
5	0	0
50	0	74
Mean=5		

Group B (y)	y - M	$(y-M)^2$
4	-3.50	12.25
6	-1.50	2.25
8	0.50	0.25
2	-5.50	30.25
9	1.50	2.25
1	-6.5	42.25
12	4.50	20.25
13	5.50	30.25
10	2.50	6.25
10	2.50	6.25
75	0	152.50
mean=7.5		

Null Hypothesis - There is no significant difference in the mean score of both groups.

$n_1 = 10$

Std. Deviation

$$S_1 = \sqrt{\frac{\Sigma(x-M)^2}{n-1}} = \sqrt{\frac{74}{9}}$$

$$= \sqrt{8.22} = 2.87$$

$n_2 = 10$

Std. Deviation

$$S_2 = \sqrt{\frac{\Sigma(y-m)^2}{n-1}} = \sqrt{\frac{152.5}{9}} = \sqrt{16.94}$$

$$= 4.12$$

$$t = \frac{M_1 \sim M_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{5-7.5}{\sqrt{\frac{8.22}{10} + \frac{16.94}{10}}} = \frac{-2.5}{\sqrt{2.51}} = -1.58$$

As the t-value is -1.58, less than the critical value at both level of significance, hence Null Hypothesis is accepted.